Electronic Corpora in Translation Studies

(National Seminar on **Knowledge Text Translation Scenario Assamese)**

Date: 1 & 2March, 2011.
Venue: Department of Cultural Studies,
Tezpur University.

Dr Biswajit Das
Department of Assamese
Nowgong Girls' College, Nagaon.
arhnk@rediffmail.com

Written and spoken translations have played a crucial role in intra-human communication in human history. But as an academic subject the translation studies has evolved only in the last fifty years.

Corpus translation studies, a newly adopted mode in Translation Studies are a growing result of the information age which facilitates the process of storing, retrieving and manipulating information. The corpus based electronics uses the rules of the source and target languages. It translates the meanings with the help of dictionaries and understanding the grammar of the source and target languages. The source language text is parsed into its parts and grammatical rules such as noun, pronoun, adjective, subject or predicate is attached to each word in the sentence. This step produces word-by-word substitution because of the use of sentence context. It assures following of the grammar of the target language. After the sentences are broken into its grammatical constituents, the words are recorded applying the rules. This step ensures that incoming sentence is altered to produce grammatically correct sentence in the target language. Finally contraction or inflection words are added in the target language. And then we may get a better output.

We can say that the corpus based machine translation is an attempt to automate all or part of the process of translating one human language to another.

Before we go to details of the study we should know about the corpus and corpus linguistics.

What is Corpus and corpus linguistics?

A **corpus** is a collection of spoken or written texts (the Latin word corpus means "body", hence a corpus is anybody of text) stored and processed mainly on computers. According to the '*The Dictionary of Linguistics and Phonetics' c*orpus is "a collection of language material which may be not just written texts, but also recorded scripts" (Crystal, 2003, p. 10). Corpus is mainly used to describe languages or to validate linguistic hypothesis. And thus, Corpus Linguistics is a branch of linguistics that uses a large collection of natural texts known as corpus for analysis. It is a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm, or a combination of these.

The additional feature of corpus is machine-readable. Machine-readable corpora possess the following advantages over written or spoken formats:

- They can be searched and manipulated.
- They can easily be enriched with extra information.

We will discuss here mainly on translation corpora, not the other uses of electronic corpora presently going on.

Monolingual and Multilingual corpora:

Monolingual corpora consists text with single language and Multi-lingual corpora is that contains texts with two or more languages.

The new waves of study on translation, multilingual corpora are playing an increasingly prominent role. Multilingual corpora have a special importance to contrastive linguistics since they deal with comparing the naturally occurring structures or patterns of two or more languages analyzing the texts produced by those languages.

Multilingual corpora subdivided into Parallel and comparable corpora.

Comparable Corpora

Comparable corpora are a collection of texts in single language together with the texts translated into the same language. Although comparable corpora don't have any roles in areas like translator training, materials writing or machine translation.

Parallel Corpora

Most of the latest research in translation knowledge acquisition is based on parallel corpora. Parallel corpus includes source language texts together with their translations and used to obtain information about the translational behavior of language-pairs, to investigate the relationship between lexical or structural equivalences in source text and target text.

This type of corpora can be applied in areas, like translator training, improving machine translation systems and material writing.

History of Text Corpus Generation:

In 1961, two linguists, namely, Nelson Francis and Henry Kucera, of the Brown University, USA initiated an attempt to develop a text corpus generated in electronic form. The Brown Corpus was made off with 500 test samples and each one has 2000 words. After that, The Lancaster-Oslo/Bergen corpus was a mutual collaborative work between the University of Lancaster, the University of Oslo, and the Norwegian Computing Centre for the Humanities, Bergen in 1978. The LOB Corpus contains one million word collections of present day samples of British English. The process of corpora is running on. The Australian Corpus of English, The Corpus of New Zealand English, The Freiburg –LOB Corpus of British English (1991), The British National Corpus, The American National Corpus, the Bank of English, The Croatian National Corpus, The English-Norwegian Parallel Corpus, The International Corpus of English etc is the example of Corpus history. With this generation the bilingual and multilingual corpora in several languages are growing up day by day. The result is no more problem for doing all kinds of cross-linguistic studies and research.

At present, there are many machine translation system successfully running and using mainly for technical documentation. ***Systran***, a well known system, which is currently, supports 52 language and 20 specialized domains. It integrates multilingual functionalities in information processing and exchanges, for applications such as ecommerce, Content Management, databases, email, Instant Messaging, SMS, etc.

Corpus in India:

In India, The first corpus developed in Indian languages was **Kolhapur Corpus of Indian English**, and this was started at individual level. Prof. S.V. Shastri and his team of Shivaji University, Kolhapur designed the corpora in 1988. It contains approximately one million words in Indian English drawn from materials published in the year 1978.

After a decade back the second corpus is developed with financial assistance of Government of India. The Department of Electronics, Govt. of India started the projects. "*In 1991, under the Technology Development for Indian Languages (TDIL) programme, it was decided that machine-readable corpora of nearly 10 million words would be developed within three years for all Indian national languages. Software for POS tagging, frequency count, spell-checkers, morphological processing, etc. would also be developed for Indian languages using the corpora. Indian Institute of Technology, Kanpur was entrusted to develop tools for language processing and machine-aided translation system from English to Indian languages*."(DASH, NILADRI SEKHAR: **LANGUAGE CORPORA: PRESENT INDIAN NEED)**

The project was initiated and worked out with the following languages:

Part	languages	Agency	Started	Closed	Word
I	English, Hindi, Punjabi	IIT,New Delhi	1991	1994	3 million
II	Telugu, Kannada, Tamil, Malayalam	CIIL, Mysore	1991	1994	3 million
III	Marathi, Gujarati	DC, Pune	1991	1994	3 million
IV	Assamese, Bangla, Oriya	IIALS, Bhubaneswar	1991	1994	3 million
V	Sanskrit	SSU, Varanasi	1991	1994	3 million
VI	Urdu, Sindhi, Kashmiri	AMU,Aligarh	1991	1994	3 million

Unfortunately, the project could not go to the end as they planned.

But the computer scientist and linguist decided to go with corpus linguistics and start developing corpora and using them in various technology developments and linguistic studies in Indian languages. The IIT's and the Universities of India take the responsibility to develop corpora in Indian Languages.

In India, there are many machine translation systems being developed by various organizations including A*nglaBharati* by *IIT Kanpur*, *MaTra* by *NCST, Anusaaraka* by *IIIT Hyderabad*, Mantra by C-DAC and *Anubad* by *Jadavpur University, Calcutta.*

Corpus in Assam:

IIT, Guwahati has taken a project to develop corpora in Assamese and Manipuri Language. IIT Guwahati has already developed a corpus dictionary in Assamese. The Computer Science Department of Guwahati University is also involved in multilingual translation corpus with Assamese, Boro and English Language. The Multilingual project was started in 2009. One Digital Assamese dictionary is also available in Market.

What type of translation corpora we needed? The answer might be- The translation should not be 'word to word' replacement of the texts form one language to another. The translation of target language should be grammatically correct and conceptually acceptable. The output should be nearest to the source language both in sense and context.

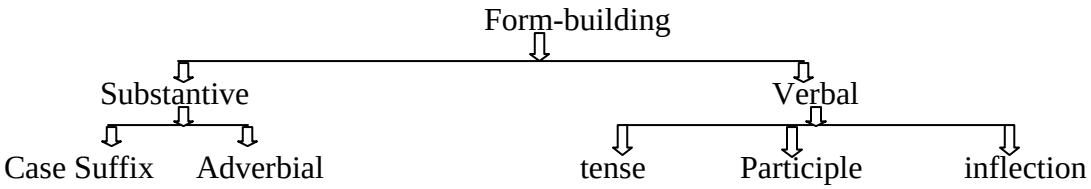
But this is not really an easy task to solve.

What kinds of problems may arise?

Assamese language represents Indo-Aryan Family. The word formation system and the sentence structure system are different from the English language.

In Assamese, the *Pada* are use in a sentence. The *Pada* is '*Sabda* plus the grammatical suffixes'. (Taraporewala, 1962) So, it is also equally important to collect the affixes using in sentence. After that, there could be some analysis, in which word the affixes will be added, and where. There are two types of affixes use in Assamese. These are- Prefixes (Upasarga) and Suffixes (Parasarga).

In the same way, the form-building are also as important as the world-building. The form building process is-



These are basic things which will need to be analysis for preparing a translate corpora. The machine translation is based on theory. Therefore, this type of category is needed to be theorized.

The word formation process is just an example only. In the same way the other grammatical or morphological forms are also need to be analysis. This means the translation systems have required details about grammatical, morphological variants.

Grammatical analysis helps us to explore and explicate how languages are structured. This type of theorized study has not been done till date to prepare a corpora in Assamese. But Dr **Shikar Sarmah** of Computer Science, Gauhati University and his team, is doing a wonderful work for

preparing of this multilingual corpus. At present **they have already stored 1.5 million Assamese words drawn from printed materials covering different fields since 1972.**

Phrases and Idioms

Idioms and phrases are particular modes of expression using in natural language consist with two or more words related to socio cultural phenomena. It is not easy to translate to another language, even for human translator. The word by word translation gives different meaning.

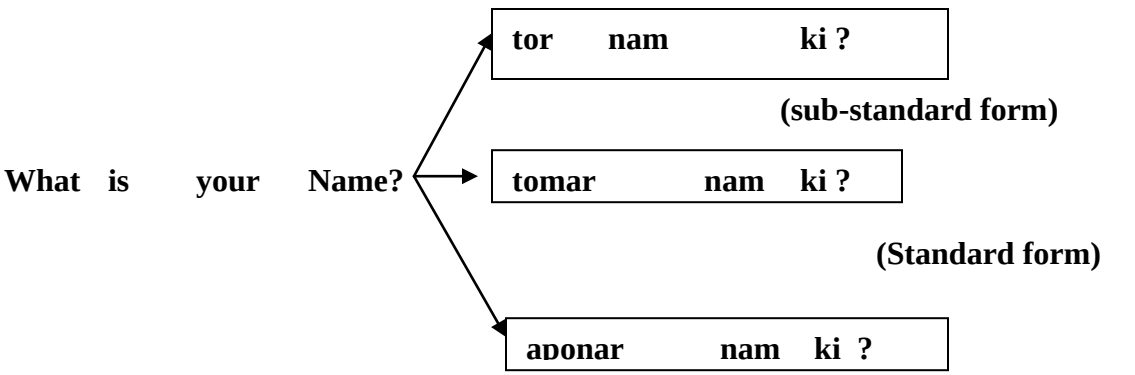
There are some other problems like- polysemy, homonymy, synonyms, Metaphors and symbols etc. Polysemy is the ambiguity of an individual word or phrases have two or more meanings or may be use in different context. Which one meaning will be taken? This is really a tough task to take the right meaning. Homonymy is that situation where a word has same pronunciation or same spelling but difference in meaning. Metaphors and Symbols are depending on the underlying culture and history, which often cannot be translated.

We can mention here an example of polysemy. Assamese and English language has different pronouns.

	Assamese(sl)	English(sl)	Assamese(pl)	English(pl)
Ist	mai	I	ami	we
2nd	tai	you	tahat	you
2nd	Tumi	you	tomalok	you
2nd	apuni	you	aponalok	you
3rd	xi/tai	he/she	xihat	they

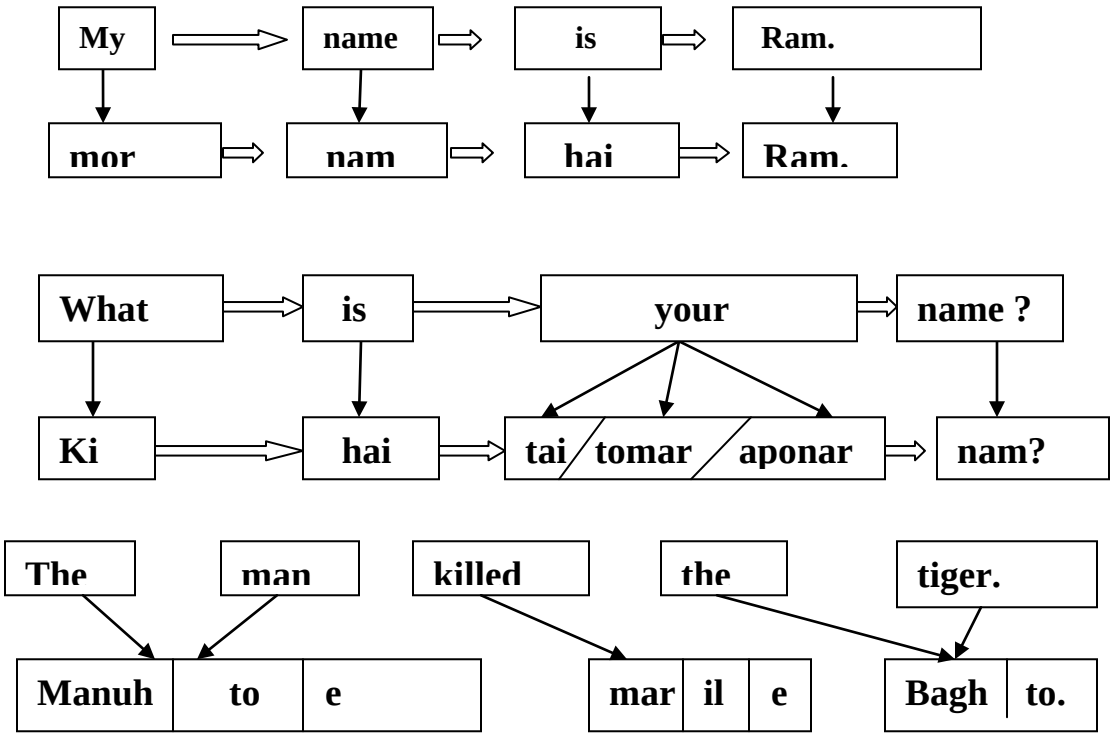
In Assamese, There are three forms in 2nd person; But English has only one word. If Assamese is target language and English is the source language, then some problems may arise. And that problem is like as –

Source Language: (English) **Target language (Assamese)**

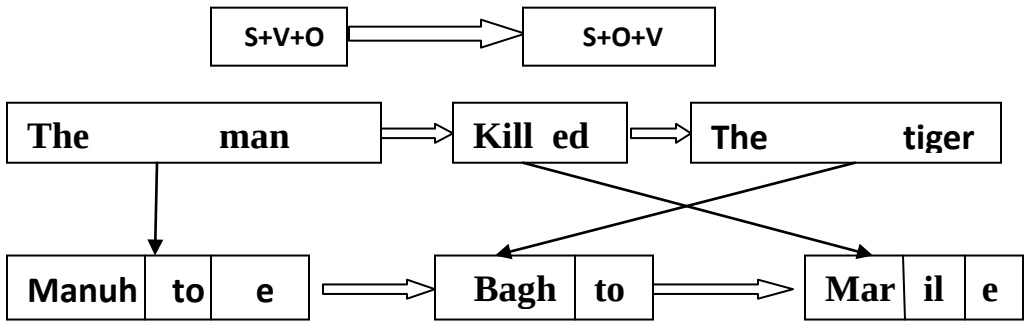


Regarding this matter, what word will accept from target language to translate the source language word '**you**'? The answer might be, a prepared theory or software accept the right word.

Sentence is the basic unit of a language. The sentence structure of the Assamese is different from English. The English language structure is S+V+O. On the other hand, the sentence structure of Assamese is S+O+V. What will be happened if machine translates a sentence word by word? That will be –



It may be possible to change the structure of target language with the help of software. Then these sentences seem to be-

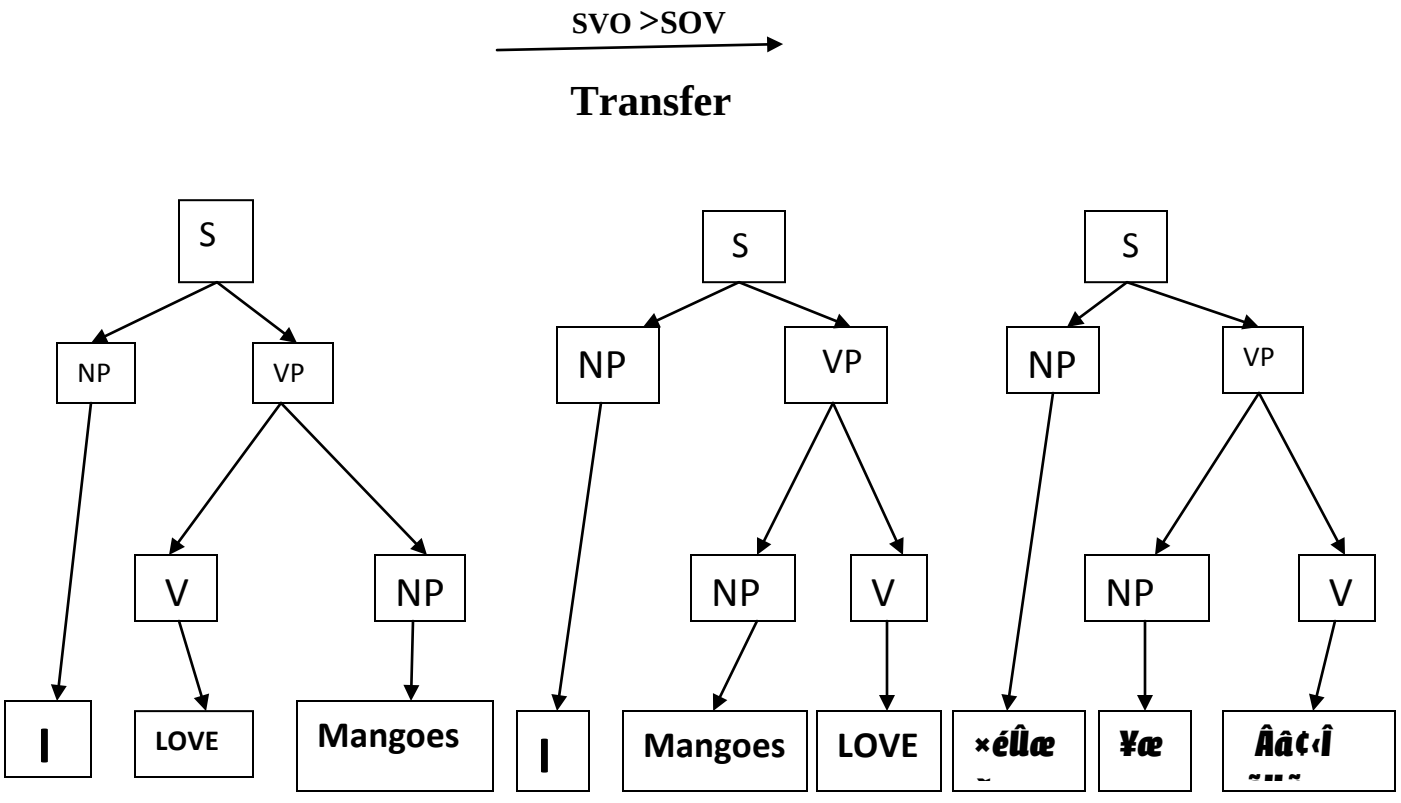


'**mor nam ram hai'** and '**tor/tomar nam ki hai'** both the sentences are not correct. There is another problem for the translation software developers. That is,

Sometimes in Assamese language the verb has not been seen as always seen in English. The correct sentences are as-

mor	**nam**	**ram.**
tor/ tomar	**nam**	**ki?**

Pushpak Bhattacharya, Department of Computer Science and Engineering, Indian Institute of Technology, Mumbai wrote a article on **'Machine Translation, Language Divergence and Lexical Resources**' and he mentioned that how does the transfer can happen in machine translation**.** Bhattacharya saw this example with English-Hindi language-



These are only simple sentence example. The natural language translation is really a tough task. It will be tougher when we go to the complex or compound sentence.

The Indo Aryan Languages are to some extent similar to each other. Bhattacharya's example is similar to Assamese. In present time, many machine translation studies are going on in Indian languages. The corpus based machine translation on Bangla and Oriya language is also going on. Knowledge transmission between the researchers of different similar Indian languages can help to build up a new technology for translation.

Idioms and Collocations are also difficult challenge in translation, as their meaning can't be derived from its Constituents. Another problem is, Phrasal verbs have different meanings in different contexts. Definitely, question would be raised, how machine can translate the meaning according to their context.

Language is a social institution. Social behavior is reflected through the language. Different social groups have different social behavior. Translation work demands, this social behavior should reflect in target language as described in source language. Is it possible in machine translation?

These types of questions have been already raised. The machine translation itself is a new conception. Of course, it is unrealistic to expect for an unbelievable result immediately. But it is certain that tools will play an important role in translation.

Conclusion:

Translation work requires various types of knowledge. Translating Natural language text is actually understanding of the language. This includes the knowledge of source and target language, their syntactic structure, word–to-word correspondence, knowledge about the domain, common sense, social conventions, etc. A human translator uses a number of knowledge sources, a wide variety of context and background information to arrive at a target language text as close as possible to the original source language text.

A multilingual country like India desperately needs an electronic corpus for translation. This translation corpus can help many ways. Many Indian languages are already endangered. This corpus translation may help the languages to survive. Not only this, this corpus also helps to spread the knowledge in grassroots level.

We can hope to see multilingual translation corpora in Indian languages.

Note:

Systran: A leading supplier of language translation Software.

References:

1. Dash, Niladri Sekhar: Language Corpora, Mittal Publication, 2009.
2. Crystal, David: The Dictionary of Linguistics and Phonetics, Oxford, 2003.
3. Taraporewala, I.J.S.: The Elements of Science in Language, Culcutta University.
4. Kakati, Banikanta: Assamese, Its Formation and Development,
5. Bhattacharyya, Pushpak Bhattacharyya: Machine Translation, Language Divergence and Lexical Resources
6. Dash, Niladri Sekhar: Machine Translation, The present Indian Need
7. Dash, Niladri Sekhar: Corpus Linguistics : A General Introduction
8. Sinha, R.M.K. : Machine Translation : An Indian Perspective.
9. Saha, Indranil; Ananthakrishnan, R; Sasikumar M: Example-Based Technique for Disambiguating Phrasal Verbs in English to Hindi Translation
10. Maria Tymoczko: Computerized Corpora and the Future of Translation Studies; Meta : journal des traducteurs / Meta: Translators' Journal, vol. 43, n° 4, 1998, p. 652-660.
11. Andrew, Chesterman: Interpreting the Meaning of Translation
12. Hutchins, W. John: Machine translation: half a century of research and use; http://ourworld.compuserve.com/homepages /WJHutchins
13. Das, Pradeep Kumar: The present status and future prospects of Computational linguistics in India.
14. CDAC URL: http://www.cdac.org.in
15. Gey, Fredric C.: Prospects for Machine Translation of the Tamil Language. gey@ucdata.berkeley.edu
16. Corpora URL: http://www.iiit.net